

# Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/FR05/000564

International filing date: 09 March 2005 (09.03.2005)

Document type: Certified copy of priority document

Document details: Country/Office: FR  
Number: 0403403  
Filing date: 31 March 2004 (31.03.2004)

Date of receipt at the International Bureau: 20 May 2005 (20.05.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)



World Intellectual Property Organization (WIPO) - Geneva, Switzerland  
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse



# BREVET D'INVENTION

CERTIFICAT D'UTILITÉ - CERTIFICAT D'ADDITION

## COPIE OFFICIELLE

Le Directeur général de l'Institut national de la propriété industrielle certifie que le document ci-annexé est la copie certifiée conforme d'une demande de titre de propriété industrielle déposée à l'Institut.

Fait à Paris, le 02 MARS 2005

Pour le Directeur général de l'Institut  
national de la propriété industrielle  
Le Chef du Département des brevets

Martine PLANCHE

INSTITUT  
NATIONAL DE  
LA PROPRIÉTÉ  
INDUSTRIELLE

SIEGE  
26 bis, rue de Saint-Petersbourg  
75800 PARIS cedex 08  
Téléphone : 33 (0)1 53 04 53 04  
Télécopie : 33 (0)1 53 04 45 23  
www.inpi.fr





26 bis, rue de Saint Pétersbourg - 75800 Paris Cedex 08

Pour vous informer : INPI DIRECT

 0 825 83 85 87  
 0,15 € TTC/mn

Télécopie : 33 (0)1 53 04 52 65

Réservé à l'INPI

**BREVET D'INVENTION**  
**CERTIFICAT D'UTILITÉ**

Code de la propriété intellectuelle - Livre VI



N° 11354\*03

**REQUÊTE EN DÉLIVRANCE**

page 1/2



Cet imprimé est à remplir lisiblement à l'encre noire

DB 540 @ W / 030103

REMISE DES PIÈCES DATE <b>31 MARS 2004</b> LIEU <b>75 INPI PARIS 34 SP</b> N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI DATE DE DÉPÔT ATTRIBUÉE PAR L'INPI <b>0403403</b> <b>31 MARS 2004</b>		<input checked="" type="checkbox"/> <b>NOM ET ADRESSE DU DEMANDEUR OU DU MANDATAIRE À QUI LA CORRESPONDANCE DOIT ÊTRE ADRESSÉE</b> " CABINET LAVOIX 2 Place d'Estienne d'Orves 75441 PARIS CEDEX 09 "	
<b>Vos références pour ce dossier</b> (facultatif) BFF 04P0012			
<b>Confirmation d'un dépôt par télécopie</b>		<input type="checkbox"/> N° attribué par l'INPI à la télécopie	
<b>2 NATURE DE LA DEMANDE</b>		<b>Cochez l'une des 4 cases suivantes</b>	
Demande de brevet		<input checked="" type="checkbox"/>	
Demande de certificat d'utilité		<input type="checkbox"/>	
Demande divisionnaire		<input type="checkbox"/>	
<i>Demande de brevet initiale</i> <i>ou demande de certificat d'utilité initiale</i>		N°	Date
		N°	Date
Transformation d'une demande de brevet européen <i>Demande de brevet initiale</i>		<input type="checkbox"/>	
		N°	Date
<b>3 TITRE DE L'INVENTION (200 caractères ou espaces maximum)</b> Procédé et système améliorés de conversion d'un signal vocal.			
<b>4 DÉCLARATION DE PRIORITÉ OU REQUÊTE DU BÉNÉFICE DE LA DATE DE DÉPÔT D'UNE DEMANDE ANTÉRIEURE FRANÇAISE</b>		Pays ou organisation Date <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> N° Pays ou organisation Date <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> N° Pays ou organisation Date <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> N° <input type="checkbox"/> S'il y a d'autres priorités, cochez la case et utilisez l'imprimé «Suite»	
<b>5 DEMANDEUR (Cochez l'une des 2 cases)</b>		<input checked="" type="checkbox"/> <b>Personne morale</b> <input type="checkbox"/> <b>Personne physique</b>	
Nom ou dénomination sociale		FRANCE TELECOM	
Prénoms			
Forme juridique		Société Anonyme	
N° SIREN		<input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/>	
Code APE-NAF		<input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/>	
Domicile ou siège	Rue	6, Place d'Alleray	
	Code postal et ville	75 011 5 PARIS	
	Pays	FRANCE	
Nationalité		Française	
N° de téléphone (facultatif)		N° de télécopie (facultatif)	
Adresse électronique (facultatif)			
<input type="checkbox"/> S'il y a plus d'un demandeur, cochez la case et utilisez l'imprimé «Suite»			

Remplir impérativement la 2<sup>ème</sup> page



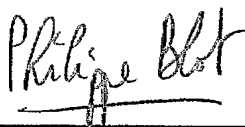
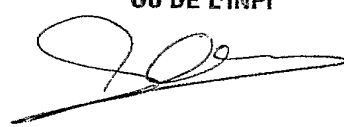
# BREVET D'INVENTION CERTIFICAT D'UTILITÉ

REQUÊTE EN DÉLIVRANCE  
page 2/2

BR2

REMISE DES RECHERCHES	31 MARS 2004
DATE	75 INPI PARIS 34 SP
LIEU	0403403
N° D'ENREGISTREMENT	
NATIONAL ATTRIBUÉ PAR L'INPI	

DB 540 W / 191203

<b>6 MANDATAIRE (s'il y a lieu)</b>			
Nom			
Prénom			
Cabinet ou Société		CABINET LAVOIX	
Nationalité			
N° de pouvoir permanent et/ou de lien contractuel			
Adresse	Rue	2 Place d'Estienne d'Orves	
	Code postal et ville	75 15 14 11 PARIS CEDEX 09	
	Pays	FRANCE	
N° de téléphone (facultatif)		01 53 20 14 20	
N° de télécopie (facultatif)		01 53 20 14 91	
Adresse électronique (facultatif)		brevets@cabinet-lavoix.com	
<b>7 INVENTEUR (S)</b>		<b>Les inventeurs sont nécessairement des personnes physiques</b>	
Les demandeurs et les inventeurs sont les mêmes personnes		<input type="checkbox"/> Oui <input checked="" type="checkbox"/> Non : Dans ce cas remplir le formulaire de Désignation d'inventeur(s)	
<b>8 RAPPORT DE RECHERCHE</b>		<b>Uniquement pour une demande de brevet (y compris division et transformation)</b>	
Établissement immédiat ou établissement différé		<input checked="" type="checkbox"/> <input type="checkbox"/> Choix à faire obligatoirement au dépôt (cf. Notice explicative Rubrique 8)	
<b>9 RÉDUCTION DU TAUX DES REDEVANCES</b>		<b>Uniquement pour les personnes physiques</b> <input type="checkbox"/> Requête pour la première fois pour cette invention (joindre un avis de non-imposition) <input type="checkbox"/> Obtenue antérieurement à ce dépôt pour cette invention (joindre une copie de la décision d'admission à l'assistance gratuite ou indiquer sa référence) : AG <input type="text"/>	
<b>10 SÉQUENCES DE NUCLEOTIDES ET/OU D'ACIDES AMINÉS</b>		<input type="checkbox"/> Cochez la case si la description contient une liste de séquences	
Le support électronique de données est joint		<input type="checkbox"/>	
La déclaration de conformité de la liste de séquences sur support papier avec le support électronique de données est jointe		<input type="checkbox"/>	
Si vous avez utilisé l'imprimé «Suite», indiquez le nombre de pages jointes			
<b>11 SIGNATURE DU DEMANDEUR OU DU MANDATAIRE</b> (Nom et qualité du signataire) Ph. BLOT N° 98-0404 		<b>VISA DE LA PRÉFECTURE OU DE L'INPI</b> 	

La présente invention concerne un procédé de conversion d'un signal vocal prononcé par un locuteur source en un signal vocal converti dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible et un système de conversion correspondant.

5 Dans le cadre d'applications de conversion de voix, telles que les services vocaux, les applications de dialogue oral homme-machine ou encore la synthèse vocale de textes, le rendu auditif est primordial et, pour obtenir une qualité acceptable, il convient de bien maîtriser les paramètres acoustiques des signaux vocaux.

10 De manière classique, les principaux paramètres acoustiques ou prosodiques modifiés lors de procédés de conversion de voix sont les paramètres relatifs à l'enveloppe spectrale, et pour les sons voisés faisant intervenir la vibration des cordes vocales, les paramètres relatifs à une structure périodique, soit la période fondamentale dont l'inverse est appelé fréquence fondamentale  
15 ou « pitch ».

Les procédés de conversion de voix classiques sont essentiellement fondés sur des modifications des caractéristiques d'enveloppe spectrale et des modifications globales des caractéristiques de fréquence fondamentale.

20 Une étude plus récente, publiée à l'occasion de la conférence EUROSPEECH 2003 sous le titre « *A new method for pitch prediction from spectral envelope and its application in voice conversion* » par Taoufik En-Najjary, Olivier Rosec and Thierry Chonavel, prévoit la possibilité d'affiner la modification des caractéristiques de fréquence fondamentale en définissant une fonction de prédiction de ces caractéristiques, en fonction de caractéristiques  
25 d'enveloppe spectrale.

Ainsi, ce procédé permet de modifier les caractéristiques d'enveloppe spectrale, et en fonction de celles-ci, de modifier les caractéristiques de fréquence fondamentale.

30 Ce procédé présente toutefois l'inconvénient important de rendre la modification des caractéristiques de fréquence fondamentale dépendantes de la modification des caractéristiques d'enveloppe spectrale. Ainsi une erreur de transformation de l'enveloppe spectrale se répercute automatiquement sur la prédiction de fréquence fondamentale.

De plus, la mise en œuvre d'un tel procédé requiert deux étapes importantes de calcul, soit la modification des caractéristiques d'enveloppe spectrale et la prédiction de la fréquence fondamentale, aboutissant ainsi à doubler la complexité du système dans son ensemble.

5 Le but de la présente invention est de résoudre ces problèmes en définissant un procédé de conversion de voix simple et plus efficace.

A cet effet, la présente invention a pour objet un procédé de conversion d'un signal vocal prononcé par un locuteur source en un signal vocal converti dont les caractéristiques acoustiques ressemblent à celles d'un locuteur  
10 cible, comprenant :

- la détermination d'au moins une fonction de transformation de caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches de celles du locuteur cible, à partir d'échantillons vocaux des locuteurs source et cible ; et

15 - la transformation de caractéristiques acoustiques du signal vocal à convertir du locuteur source, par l'application de ladite au moins une fonction de transformation,

caractérisé en ce que ladite détermination comprend la détermination d'une fonction de transformation conjointe de caractéristiques  
20 relatives à l'enveloppe spectrale et de caractéristiques relatives à la fréquence fondamentale du locuteur source et en ce que ladite transformation comprend l'application de ladite fonction de transformation conjointe.

Ainsi, le procédé de l'invention permet la modification simultanée au cours d'une seule opération des caractéristiques d'enveloppe spectrale et de  
25 fréquence fondamentale sans créer de dépendance entre celles-ci.

Suivant d'autres caractéristiques de l'invention :

- ladite détermination d'une fonction de transformation conjointe comprend :

- une étape d'analyse des échantillons vocaux des locuteurs  
30 source et cible regroupés en trames pour obtenir, pour chaque trame d'échantillons d'un locuteur, des informations relatives à l'enveloppe spectrale et à la fréquence fondamentale ;

- une étape de concaténation des informations relatives à l'enveloppe spectrale et à la fréquence fondamentale pour chacun des locuteurs source et cible ;
- une étape de détermination d'un modèle représentant des caractéristiques acoustiques communes des échantillons vocaux du locuteur source et du locuteur cible ; et
- une étape de détermination, à partir de ce modèle et des échantillons vocaux, de ladite fonction de transformation conjointe ;
- lesdites étapes d'analyse des échantillons vocaux des locuteurs source et cible sont adaptées pour délivrer lesdites informations relatives à l'enveloppe spectrale sous la forme de coefficients cepstraux ;
- lesdites étapes d'analyse comprennent chacune la modélisation des échantillons vocaux selon une somme d'un signal harmonique et d'un signal de bruit qui comprend :
  - une sous-étape d'estimation de la fréquence fondamentale des échantillons vocaux ;
  - une sous-étape d'analyse synchronisée de chaque trame d'échantillons sur sa fréquence fondamentale ; et
  - une sous-étape d'estimation de paramètres d'enveloppe spectrale de chaque trame d'échantillons.
- ladite étape de détermination d'un modèle correspond à la détermination d'un modèle de mélange de densités de probabilités gaussiennes ;
- ladite étape de détermination d'un modèle comprend :
  - une sous-étape de détermination d'un modèle correspondant à un mélange de densité de probabilités gaussiennes, et
  - une sous-étape d'estimation des paramètres du mélange de densités de probabilités gaussiennes à partir de l'estimation du maximum de vraisemblance entre les caractéristiques acoustiques des échantillons des locuteurs source et cible et le modèle ;
- ladite détermination d'au moins une fonction de transformation, comporte en outre une étape de normalisation de la fréquence fondamentale des trames d'échantillons des locuteurs source et cible respectivement par rapport aux moyennes des fréquences fondamentales des échantillons analysés des locuteurs source et cible ;



- le procédé comporte une étape d'alignement temporel des caractéristiques acoustiques du locuteur source avec les caractéristiques acoustiques du locuteur cible, cette étape étant réalisée avant ladite étape de détermination d'un modèle ;

5                   - le procédé comporte une étape de séparation dans les échantillons vocaux du locuteur source et du locuteur cible, des trames à caractère voisé et des trames à caractère non voisé, ladite détermination d'une fonction de transformation conjointe des caractéristiques relatives à l'enveloppe spectrale et à la fréquence fondamentale étant réalisée uniquement à partir  
10 desdites trames voisées et le procédé comportant une détermination d'une fonction de transformation des seules caractéristiques d'enveloppe spectrale uniquement à partir desdites trames non voisées ;

- ladite détermination d'au moins une fonction de transformation comprend uniquement ladite étape de détermination d'une fonction de  
15 transformation conjointe ;

- ladite détermination d'une fonction de transformation conjointe est réalisée à partir d'un estimateur de la réalisation des caractéristiques acoustiques du locuteur cible sachant les caractéristiques acoustiques du locuteur source ;

- ledit estimateur est formé de l'espérance conditionnelle de la  
20 réalisation des caractéristiques acoustiques du locuteur cible sachant la réalisation des caractéristiques acoustiques du locuteur source ;

- ladite transformation de caractéristiques acoustiques du signal vocal à convertir, comporte :

- une étape d'analyse de ce signal vocal, regroupé en trames  
25 pour obtenir, pour chaque trame d'échantillons, des informations relatives à l'enveloppe spectrale et à la fréquence fondamentale ;

- une étape de formatage des informations acoustiques relatives à l'enveloppe spectrale et à la fréquence fondamentale du signal vocal à convertir ; et

30                   - une étape de transformation des informations acoustiques formatées du signal vocal à convertir à l'aide de ladite fonction de transformation conjointe ;

- le procédé comporte une étape de séparation, dans ledit signal vocal à convertir, des trames voisées et des trames non voisées, ladite étape de transformation comprenant :

- une sous-étape d'application de ladite fonction de transformation conjointe aux seules trames voisées dudit signal à convertir ; et

- une sous-étape d'application de ladite fonction de transformation des seules caractéristiques d'enveloppe spectrale auxdites trames non voisées dudit signal à convertir ;

- ladite étape de transformation comprend l'application de ladite fonction de transformation conjointe aux caractéristiques acoustiques de toutes les trames dudit signal vocal à convertir ;

- le procédé comporte en outre une étape de synthèse permettant de former un signal vocal converti à partir des dites informations acoustiques transformées.

L'invention a également pour objet un système de conversion d'un signal vocal prononcé par un locuteur source en un signal vocal converti dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible, comprenant :

- des moyens de détermination d'au moins une fonction de transformation des caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches du locuteur cible, à partir d'échantillons vocaux prononcés par les locuteurs source et cible : et

- des moyens de transformation des caractéristiques acoustiques du signal vocal à convertir du locuteur source par l'application de ladite au moins une fonction de transformation,

caractérisé en ce que lesdits moyens de détermination d'au moins une fonction de transformation, comprennent une unité de détermination d'une fonction de transformation conjointe de caractéristiques relatives à l'enveloppe spectrale et de caractéristiques relatives à la fréquence fondamentale du locuteur source et en ce que lesdits moyens de transformation comportent des moyens d'application de ladite fonction de transformation conjointe.

Selon d'autres caractéristiques de ce système :

- il comporte en outre :

- des moyens d'analyse du signal vocal à convertir, adaptés pour délivrer en sortie des informations relatives à l'enveloppe spectrale et à la fréquence fondamentale du signal vocal à convertir ; et

5       - des moyens de synthèse permettant de former un signal vocal converti à partir au moins desdites informations d'enveloppe spectrale et de fréquence fondamentale transformées simultanément ;

10       - lesdits moyens de détermination d'au moins une fonction de transformation de caractéristiques acoustiques comportent en outre une unité de détermination d'une fonction de transformation de l'enveloppe spectrale des trames non voisées, ladite unité de détermination de la fonction de transformation conjointe étant adaptée pour la détermination de la fonction de transformation conjointe uniquement pour les trames voisées.

15       L'invention sera mieux comprise à la lecture de la description qui va suivre, donnée uniquement à titre d'exemple et faite en se référant aux dessins annexés, sur lesquels :

- les Figs. 1A et 1B forment un organigramme général d'un premier mode de réalisation du procédé de l'invention ;

- les Figs. 2A et 2B forment un organigramme général d'un second mode de réalisation du procédé de l'invention ;

20       - la Fig. 3 est un graphique représentant un relevé expérimental des performances du procédé de l'invention ; et

- la Fig. 4 est un schéma synoptique d'un système mettant en œuvre un procédé selon l'invention.

25       La conversion de voix consiste à modifier le signal vocal d'un locuteur de référence appelé locuteur source, de telle sorte que le signal produit semble avoir été prononcé par un autre locuteur, nommé locuteur cible.

30       Un tel procédé comporte tout d'abord la détermination de fonctions de transformation de caractéristiques acoustiques ou prosodiques des signaux vocaux du locuteur source en caractéristiques acoustiques proches de celles des signaux vocaux du locuteur cible, à partir d'échantillons vocaux prononcés par le locuteur source et le locuteur cible.

Plus particulièrement, la détermination 1 de fonctions de transformation est réalisée sur des bases de données d'échantillons vocaux

correspondant à la réalisation acoustique de mêmes séquences phonétiques prononcées respectivement par les locuteurs source et cible.

Cette détermination est désignée sur la figure 1A par la référence numérique générale 1 et est également couramment appelée « apprentissage ».

5 Le procédé comporte ensuite une transformation des caractéristiques acoustiques d'un signal vocal à convertir prononcé par le locuteur source à l'aide de la ou des fonctions déterminées précédemment. Cette transformation est désignée par la référence numérique générale 2 sur la figure 1B.

10 Le procédé débute par des étapes 4X et 4Y d'analyse des échantillons vocaux prononcés respectivement par les locuteurs source et cible. Ces étapes permettent de regrouper les échantillons par trames, afin d'obtenir pour chaque trame d'échantillons, des informations relatives à l'enveloppe spectrale et des informations relatives à la fréquence fondamentale.

15 Dans le mode de réalisation décrit, les étapes 4X et 4Y d'analyse sont fondées sur l'utilisation d'un modèle de signal sonore sous la forme d'une somme d'un signal harmonique avec un signal de bruit selon un modèle communément appelé "HNM" (en anglais : Harmonic plus Noise Model).

20 Le modèle HNM comprend la modélisation de chaque trame de signal vocal en une partie harmonique représentant la composante périodique du signal, constituée d'une somme de L sinusoïdes harmoniques d'amplitude  $A_i$  et de phase  $\phi_i$ , et d'une partie bruitée représentant le bruit de friction et la variation de l'excitation glottale.

On peut ainsi écrire :

$$s(n)=h(n)+b(n)$$

25 avec 
$$h(n)=\sum_{i=1}^L A_i(n)\cos(\phi_i(n))$$

Le terme  $h(n)$  représente donc l'approximation harmonique du signal  $s(n)$ .

En outre, le mode de réalisation décrit est fondé sur une représentation de l'enveloppe spectrale par le cepstre discret.

30 Les étapes 4X et 4Y comportent des sous-étapes 8X et 8Y d'estimation pour chaque trame, de la fréquence fondamentale, par exemple au moyen d'une méthode d'autocorrélation.

Les sous-étapes 8X et 8Y sont chacune suivies d'une sous-étape 10X et 10Y d'analyse synchronisée de chaque trame sur sa fréquence fondamentale, qui permet d'estimer les paramètres de la partie harmonique ainsi que les paramètres du bruit du signal et notamment la fréquence maximale de  
 5 voisement. En variante, cette fréquence peut être fixée arbitrairement ou être estimée par d'autres moyens connus.

Dans le mode de réalisation décrit, cette analyse synchronisée correspond à la détermination des paramètres des harmoniques par minimisation d'un critère de moindres carrés pondérés entre le signal complet et sa  
 10 décomposition harmonique correspondant dans le mode de réalisation décrit, au signal de bruit estimé. Le critère noté E est égal à :

$$E = \sum_{n=-T_i}^{T_i} w^2(n)(s(n)-h(n))^2$$

Dans cette équation,  $w(n)$  est la fenêtre d'analyse et  $T_i$  est la période fondamentale de la trame courante.

15 Ainsi, la fenêtre d'analyse est centrée autour de la marque de la période fondamentale et a pour durée deux fois cette période.

En variante, ces analyses sont faites de manière asynchrone avec un pas fixe d'analyse et une fenêtre de taille fixe.

Les étapes 4X et 4Y d'analyse comportent enfin des sous-étapes 12X et 12Y d'estimation des paramètres de l'enveloppe spectrale des signaux en  
 20 utilisant par exemple une méthode de cepstre discret régularisé et une transformation en échelle de Bark pour reproduire le plus fidèlement possible les propriétés de l'oreille humaine.

Ainsi, les étapes 4X et 4Y d'analyse délivrent respectivement pour les  
 25 échantillons vocaux prononcés par les locuteurs source et cible, pour chaque trame de rang  $n$  d'échantillons des signaux de parole, un scalaire noté  $F_n$  représentant la fréquence fondamentale et un vecteur noté  $c_n$  comprenant des informations d'enveloppe spectrale sous la forme d'une séquence de coefficients cepstraux.

30 Le mode de calcul des coefficients cepstraux correspond à un mode opératoire connu de l'état de la technique et, pour cette raison, ne sera pas décrit plus en détail.

- Avantageusement, les étapes 4X et 4Y d'analyse sont suivies chacune par une étape 14 X et 14Y de normalisation de la valeur de la fréquence fondamentale de chaque trame par rapport respectivement aux fréquences fondamentales des locuteurs source et cible afin de remplacer, pour chaque
- 5 trame d'échantillons vocaux, la valeur de la fréquence fondamentale par une valeur de fréquence fondamentale normalisée selon la formule suivante :

$$g = F_{\log} = \log \left( \frac{F_o}{F_o^{\text{moy}}} \right)$$

- Dans cette formule,  $F_o^{\text{moy}}$  correspond aux moyennes des valeurs des fréquences fondamentales sur chaque base de données analysée, soit sur la
- 10 base de données d'échantillons vocaux du locuteur source et du locuteur cible.

- Cette normalisation permet de modifier, pour chaque locuteur, l'échelle de variations des scalaires de fréquence fondamentale afin de la rendre cohérente avec l'échelle des variations des coefficients cepstraux. Pour chaque trame n, on note  $g_x(n)$  la fréquence fondamentale normalisée pour le locuteur
- 15 source et  $g_y(n)$  celle du locuteur cible.

Le procédé de l'invention comporte ensuite des étapes 16X et 16Y de concaténation pour chaque locuteur source et cible, des informations d'enveloppe spectrale et de fréquence fondamentale sous la forme d'un unique vecteur.

- 20 Ainsi, l'étape 16X permet de définir pour chaque trame n un vecteur noté  $x_n$  regroupant les coefficients cepstraux  $c_x(n)$  et la fréquence fondamentale normalisée  $g_x(n)$  selon l'équation suivante :

$$x_n = \left[ c_x^T(n), g_x(n) \right]^T$$

Dans cette équation, T désigne l'opérateur de transposition.

- 25 De manière similaire, l'étape 16Y permet de former pour chaque trame n, un vecteur  $y_n$  reprenant les coefficients cepstraux  $c_y(n)$  et la fréquence fondamentale normalisée  $g_y(n)$  selon l'équation suivante :

$$y_n = \left[ c_y^T(n), g_y(n) \right]^T$$

Les étapes 16 X et 16Y sont suivies d'une étape 18 d'alignement entre le vecteur source  $x_n$  et le vecteur cible  $y_n$ , de manière à former un appariement entre ces vecteurs obtenu par un algorithme classique d'alignement temporel dynamique dit « DTW » (en anglais : Dynamic Time Warping).

- 5 En variante, l'étape 18 d'alignement est mise en œuvre uniquement à partir des coefficients cepstraux sans utiliser les informations de fréquence fondamentale.

L'étape 18 d'alignement délivre donc un vecteur couple formé de couples de coefficients cepstraux et d'informations de fréquence fondamentale  
10 des locuteurs source et cible, alignés temporellement.

L'étape 18 d'alignement est suivie d'une étape 20 de détermination d'un modèle représentant les caractéristiques acoustiques communes du locuteur source et du locuteur cible à partir des informations d'enveloppe spectrale et de fréquence fondamentale de tous les échantillons analysés.

- 15 Dans le mode de réalisation décrit, il s'agit d'un modèle probabiliste des caractéristiques acoustiques du locuteur cible et du locuteur source, selon un modèle de mélange de densités de probabilités gaussiennes, couramment noté "GMM", dont les paramètres sont estimés à partir des vecteurs source et cible contenant, pour chaque locuteur, la fréquence fondamentale normalisée et le  
20 cepstre discret.

De manière classique, la densité de probabilité d'une variable aléatoire notée de manière générale  $p(z)$ , suivant un modèle de mélange de densités gaussiennes GMM s'écrit mathématiquement de la manière suivante :

$$p(z) = \sum_{i=1}^Q \alpha_i x(z, \mu_i, \Sigma_i)$$

- 25 avec  $\sum_{i=1}^Q \alpha_i = 1, 0 \leq \alpha_i \leq 1$

Dans cette formule, Q désigne le nombre de composantes du modèle,  $N(z ; \mu_i, \Sigma_i)$  est la densité de probabilité de la loi normale de moyenne  $\mu_i$  et de matrice de covariance  $\Sigma_i$  et les coefficients  $\alpha_i$  sont les coefficients du mélange.

- Ainsi, le coefficient  $\alpha_i$  correspond à la probabilité a priori que la  
30 variable aléatoire  $z$  soit générée par la  $i^{\text{ème}}$  composante gaussienne du mélange.

De manière plus particulière, l'étape 20 de détermination du modèle comporte une sous-étape 22 de modélisation de la densité jointe  $p(z)$  des vecteurs source noté  $x$  et cible noté  $y$ , de sorte que :

$$Z_n = \begin{bmatrix} T & T \\ x_n & y_n \end{bmatrix}^T$$

5 L'étape 20 comporte ensuite une sous-étape 24 d'estimation de paramètres GMM ( $\alpha, \mu, \Sigma$ ) de la densité  $p(z)$ . Cette estimation peut être réalisée, par exemple, à l'aide d'un algorithme classique de type dit "EM" (Expectation – Maximisation), correspondant à une méthode itérative conduisant à l'obtention d'un estimateur de maximum de vraisemblance entre les données des  
10 échantillons de parole et le modèle de mélange de gaussiennes.

La détermination des paramètres initiaux du modèle GMM est obtenue à l'aide d'une technique classique de quantification vectorielle.

L'étape 20 de détermination de modèle délivre ainsi les paramètres d'un mélange de densités gaussiennes, représentatif des caractéristiques  
15 acoustiques communes et en particulier d'enveloppe spectrale et de fréquence fondamentale, des échantillons vocaux du locuteur source et du locuteur cible.

Le procédé comporte ensuite une étape 30 de détermination, à partir du modèle et des échantillons vocaux, d'une fonction conjointe de transformation de la fréquence fondamentale et de l'enveloppe spectrale fournie par le cepstre,  
20 du signal du locuteur source vers le locuteur cible.

Cette fonction de transformation est déterminée à partir d'un estimateur de la réalisation des caractéristiques acoustiques du locuteur cible étant donné les caractéristiques acoustiques du locuteur source, formé dans le mode de réalisation décrit, par l'espérance conditionnelle.

25 Pour cela, l'étape 30 comporte une sous-étape 32 de détermination de l'espérance conditionnelle des caractéristiques acoustiques du locuteur cible sachant les informations caractéristiques acoustiques du locuteur source. L'espérance conditionnelle est notée  $F(x)$  et est déterminée à partir des formules suivantes :

30 
$$F(x) = E[y | x] = \sum_{i=1}^Q h_i(x) \left[ \mu_i^y + \Sigma_i^{yx} (\Sigma_i^{xx})^{-1} (x - \mu_i^x) \right]$$



$$\text{avec } h_i(x) = \frac{\alpha N(x, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^Q \alpha N(x, \mu_j^x, \Sigma_j^{xx})}$$

$$\text{avec } \Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} \text{ et } \mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$$

Dans ces équations,  $h_i(x)$  correspond à la probabilité a posteriori que le vecteur source  $x$  soit généré par la  $i^{\text{ème}}$  composante du modèle de mélange de densités gaussiennes du modèle.

La détermination de l'espérance conditionnelle permet ainsi d'obtenir la fonction de transformation conjointe des caractéristiques d'enveloppe spectrale et de fréquence fondamentale entre le locuteur source et le locuteur cible.

Il apparaît donc que le procédé d'analyse de l'invention permet, à partir du modèle et des échantillons vocaux, d'obtenir une fonction de transformation conjointe des caractéristiques acoustiques de fréquence fondamentale et d'enveloppe spectrale.

En référence à la figure 1B, le procédé de conversion comporte ensuite la transformation 2 d'un signal vocal à convertir prononcé par le locuteur source, lequel signal à convertir peut être différent des signaux vocaux utilisés précédemment.

Cette transformation 2 débute par une étape d'analyse 36 réalisée, dans le mode de réalisation décrit, à l'aide d'une décomposition selon le modèle HNM similaire à celles réalisées dans les étapes 4X et 4Y décrites précédemment. Cette étape 36 permet de délivrer des informations d'enveloppe spectrale sous la forme de coefficients cepstraux, des informations de fréquence fondamentale ainsi que des informations de phase et de fréquence maximale de voisement.

L'étape 36 est suivie d'une étape 38 de formatage des caractéristiques acoustiques du signal à convertir par normalisation de la fréquence fondamentale et concaténation avec les coefficients cepstraux afin de former un unique vecteur.

Cet unique vecteur est utilisé lors d'une étape 40 de transformation des caractéristiques acoustiques du signal vocal à convertir par l'application de la fonction de transformation déterminée à l'étape 30, aux coefficients cepstraux du

signal à convertir définis lors de l'étape 36, ainsi qu'aux informations de fréquence fondamentale.

A l'issue de l'étape 40, chaque trame d'échantillons du signal à convertir du locuteur source est ainsi associée à des informations d'enveloppe spectrale et de fréquence fondamentale transformées simultanément, dont les caractéristiques sont similaires à celles des échantillons du locuteur cible.

Le procédé comporte ensuite une étape 42 de dénormalisation des informations de fréquence fondamentale transformées.

Cette étape 42 permet de ramener les informations de fréquence fondamentale transformées sur une échelle propre au locuteur cible selon l'équation suivante :

$$F_o[F(x)] = F_o^{moy}(y) \cdot e^{F[g_x(n)]}$$

Dans cette équation  $F_o[F(x)]$  correspond à la fréquence fondamentale transformée dénormalisée,  $F_o^{moy}(y)$  à la moyenne des valeurs des fréquences fondamentales du locuteur cible et  $F[g_x(n)]$  à la transformée de la fréquence fondamentale normalisée du locuteur source.

De manière classique, le procédé de conversion comporte ensuite une étape 44 de synthèse du signal de sortie réalisée, dans l'exemple décrit, par une synthèse de type HNM qui délivre directement le signal vocal converti à partir des informations d'enveloppe spectrale et de fréquence fondamentale transformées délivrées par l'étape 40 et des informations de phase et de fréquence maximale de voisement délivrées par l'étape 36.

Le procédé de conversion mettant en œuvre le procédé d'analyse de l'invention permet ainsi d'obtenir une conversion de voix réalisant conjointement des modifications d'enveloppe spectrales et de fréquence fondamentale, de manière à obtenir un rendu auditif de bonne qualité.

En référence à la figure 2A, on va maintenant décrire l'organigramme général d'un second mode de réalisation du procédé de l'invention.

De même que précédemment, ce procédé comporte la détermination 1 de fonctions de transformation de caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches de celles du locuteur cible.

Cette détermination 1 débute par la mise en œuvre des étapes 4X et 4Y d'analyse des échantillons vocaux prononcés respectivement par le locuteur source et le locuteur cible.

5 Ces étapes 4X et 4Y sont fondées sur l'utilisation du modèle HNM ainsi que cela a été décrit précédemment et délivrent chacune un scalaire noté  $F(n)$  représentant la fréquence fondamentale et un vecteur noté  $c(n)$  comprenant des informations d'enveloppe spectrale sous la forme d'une séquence de coefficients cepstraux.

10 Dans ce mode de réalisation, ces étapes 4X et 4Y d'analyse sont suivies d'une étape 50 d'alignement des vecteurs de coefficients cepstraux issus de l'analyse des trames du locuteur source et des trames du locuteur cible.

Cette étape 50 est mise en œuvre par un algorithme tel que l'algorithme DTW, de manière similaire à l'étape 18 du premier mode de réalisation.

15 A l'issue de l'étape 50 d'alignement, le procédé dispose d'un vecteur couple formé de couples de coefficients cepstraux du locuteur source et du locuteur cible, alignés temporellement. Ce vecteur couple est également associé aux informations de fréquence fondamentale.

20 L'étape 50 d'alignement est suivie d'une étape 54 de séparation, dans le vecteur couple, des trames voisées et des trames non voisées.

En effet, seules les trames voisées présentent une fréquence fondamentale et un tri peut être effectué en considérant si oui ou non des informations de fréquence fondamentale existent pour chaque couple du vecteur couple.

25 Cette étape de séparation 54 permet ensuite de réaliser la détermination 56 d'une fonction de transformation conjointe des caractéristiques d'enveloppe spectrale et de fréquence fondamentale des trames voisées et la détermination 58 d'une fonction de transformation des seules caractéristiques d'enveloppe spectrale des trames non voisées.

30 La détermination 56 d'une fonction de transformation des trames voisées débute par des étapes 60X et 60Y de normalisation des informations de fréquence fondamentale respectivement pour les locuteurs source et cible.

Ces étapes 60X et 60Y sont réalisées de manière similaire aux étapes 14X et 14Y du premier mode de réalisation et aboutissent à l'obtention, pour

chaque trame voisée, de la fréquence normalisée pour le locuteur source notée  $g_x(n)$  et de celle du locuteur cible notée  $g_y(n)$ .

5 Ces étapes 60X et 60Y de normalisation sont suivies chacune d'une étape 62X et 62Y de concaténation des coefficients cepstraux  $c_x$  et  $c_y$  du locuteur source et du locuteur cible respectivement avec les fréquences normalisées  $g_x$  et  $g_y$ .

10 Ces étapes 62X et 62Y de concaténation sont réalisées de manière similaire aux étapes 16X et 16Y et permettent de délivrer un vecteur  $x_n$  contenant des informations d'enveloppe spectrale et de fréquence fondamentale pour les trames voisées du locuteur source et un vecteur  $y_n$  contenant des informations d'enveloppe spectrale et de fréquence fondamentale normalisées pour les trames voisées du locuteur cible.

15 De plus, l'alignement entre ces deux vecteurs est conservé tel qu'obtenu à l'issue de l'étape 50, les modifications survenues lors des étapes 60X et 60Y de normalisation et 62X et 62Y de concaténation étant réalisées directement à l'intérieur du vecteur délivré par l'étape 50 d'alignement.

Le procédé comporte ensuite une étape 70 de détermination d'un modèle représentant les caractéristiques communes du locuteur source et du locuteur cible.

20 A la différence de l'étape 20 décrite en référence à la figure 1A, cette étape 70 est mise en œuvre à partir des informations de fréquence fondamentale et d'enveloppe spectrale des seuls échantillons voisés analysés.

Dans ce mode de réalisation, cette étape 70 est fondée sur un modèle probabiliste selon un mélange de densité gaussienne dit GMM.

25 L'étape 70 comporte ainsi une sous-étape 72 de modélisation de la densité jointe entre les vecteurs X et Y réalisés de manière similaire à la sous-étape 22 décrite précédemment.

Cette sous-étape 72 est suivie d'une sous-étape 74 d'estimation des paramètres GMM ( $\alpha$ ,  $\mu$  et  $\Sigma$ ) de la densité  $p(z)$ .

30 De même que dans le mode de réalisation décrit précédemment, cette estimation est réalisée à l'aide d'un algorithme de type « EM » permettant l'obtention d'un estimateur de maximum de vraisemblance entre les données des échantillons de paroles et le modèle de mélange de gaussienne.

L'étape 70 délivre donc les paramètres d'un mélange de densités gaussiennes, représentatif des caractéristiques acoustiques communes d'enveloppe spectrale et de fréquence fondamentale des échantillons vocaux voisés du locuteur source et du locuteur cible.

5 L'étape 70 est suivie d'une étape 80 de détermination d'une fonction conjointe de transformation de la fréquence fondamentale et de l'enveloppe spectrale des échantillons vocaux voisés du locuteur source vers le locuteur cible.

10 Cette étape 80 est mise en œuvre de manière similaire à l'étape 30 du premier mode de réalisation et en particulier comporte également une sous-étape 82 de détermination de l'espérance conditionnelle des caractéristiques acoustiques du locuteur cible sachant les caractéristiques acoustiques du locuteur source, cette sous-étape étant mise en œuvre selon les mêmes formules que précédemment, appliquées aux seuls échantillons voisés.

15 L'étape 80 aboutit ainsi à l'obtention d'une fonction de transformation conjointe des caractéristiques d'enveloppe spectrale et de fréquence fondamentale entre le locuteur source et le locuteur cible, applicable aux trames voisées.

20 Parallèlement à la détermination 56 de cette fonction de transformation des trames voisées, la détermination 58 d'une fonction de transformation des seules caractéristiques d'enveloppe spectrale des trames non voisées est également mise en œuvre.

25 Dans le mode de réalisation décrit, la détermination 58 comporte une étape 90 de détermination d'une fonction de filtrage définie de manière globale sur les paramètres d'enveloppe spectrale, à partir des couples de trames non voisées.

Cette étape 90 est réalisée de manière classique par la détermination d'un modèle GMM ou encore de tout autre technique adaptée et connue.

30 A l'issue de la détermination 58, une fonction de transformation des caractéristiques d'enveloppe spectrale des trames non voisées est obtenue.

En référence à la figure 2B, le procédé comporte ensuite la transformation 2 des caractéristiques acoustiques d'un signal vocal à convertir.

De même que dans le mode de réalisation précédent, cette transformation 2 débute par une étape d'analyse 36 du signal vocal à convertir réalisée selon un modèle HNM et une étape 38 de formatage.

5 Ainsi que cela a été dit précédemment, ces étapes 36 et 38 permettent de délivrer, sous la forme d'un unique vecteur, les informations d'enveloppe spectrale et de fréquence fondamentale normalisée. De plus, l'étape 36 délivre des informations de phase et de fréquence maximale de voisement.

10 Dans le mode de réalisation décrit, l'étape 38 est suivie d'une étape 100 de séparation, dans le signal à convertir analysé, des trames voisées et des trames non voisées.

Cette séparation est réalisée à l'aide d'un critère fondé sur la présence d'une information de fréquence fondamentale non nulle.

15 L'étape 100 est suivie d'une étape 102 de transformation des caractéristiques acoustiques du signal vocal à convertir par l'application des fonctions de transformation déterminées lors des étapes 80 et 90.

Plus particulièrement, cette étape 102 comporte une sous-étape 104 d'application de la fonction de transformation conjointe des informations d'enveloppe spectrale et de fréquence fondamentale, déterminée à l'étape 80, aux seules trames voisées telles que séparées à l'issue de l'étape 100.

20 Parallèlement, l'étape 102 comporte une sous-étape 106 d'application de la fonction de transformation des seules informations d'enveloppe spectrale, déterminée à l'étape 90, aux seules trames non voisées telles que séparées lors de l'étape 100.

25 La sous-étape 104 délivre ainsi pour chaque trame d'échantillons voisés du signal à convertir du locuteur source, des informations d'enveloppe spectrale et de fréquence fondamentale transformées simultanément et dont les caractéristiques sont similaires à celles des échantillons voisés du locuteur cible.

30 La sous-étape 106 délivre quant à elle pour chaque trame d'échantillons non voisés du signal à convertir du locuteur source, des informations d'enveloppe spectrale transformées dont les caractéristiques sont similaires à celles des échantillons non voisés du locuteur cible.

Dans le mode de réalisation décrit, le procédé comprend en outre une étape 108 de dénormalisation des informations de fréquence fondamentale transformées, mise en œuvre sur les informations délivrées par la sous-étape

104 de transformation, d'une manière similaire à l'étape 42 décrite en référence à la figure 1B.

Le procédé de conversion comporte ensuite une étape 110 de synthèse du signal de sortie réalisée, dans l'exemple décrit, par une synthèse de type HNM qui délivre le signal vocal converti à partir des informations d'enveloppe spectrale et de fréquence fondamentale transformées ainsi que des informations de phase et de fréquence maximale de voisement pour les trames voisées et à partir des informations d'enveloppe spectrale transformées pour les trames non voisées.

Le procédé de l'invention permet donc, dans ce mode de réalisation, d'effectuer un traitement distinct sur les trames voisées et les trames non voisées, les trames voisées subissant une transformation simultanée des caractéristiques d'enveloppe spectrale et de fréquence fondamentale et les trames non voisées subissant une transformation de leurs seules caractéristiques d'enveloppe spectrale.

Un tel mode de réalisation permet une transformation plus précise que le mode de réalisation précédent tout en conservant une complexité limitée.

L'efficacité d'un procédé de conversion peut être évaluée à partir d'échantillons vocaux identiques prononcés par le locuteur source et le locuteur cible.

Ainsi, le signal vocal prononcé par le locuteur source est converti à l'aide du procédé de l'invention et la ressemblance du signal converti avec le signal prononcé par le locuteur cible est évaluée.

Par exemple, cette ressemblance est calculée sous la forme d'un rapport entre la distance acoustique séparant le signal converti du signal cible et la distance acoustique séparant le signal cible du signal source.

La figure 3 représente un graphique de résultats obtenu dans le cas d'une conversion de voix d'homme en une voix de femme, les fonctions de transformation étant obtenues à partir de bases d'apprentissage contenant chacune 5 minutes de parole échantillonnées à 16 kHz, les vecteurs cepstraux utilisés étant de taille 20 et le modèle GMM étant à 64 composantes.

Ce graphique représente en abscisse les numéros de trames et en ordonnée la fréquence en hertz du signal.

Les résultats représentés sont caractéristiques pour les trames voisées qui s'étendent approximativement des trames 20 à 85.

Sur ce graphique, la courbe Cx représente les caractéristiques de fréquence fondamentale du signal source et la courbe Cy celles du signal cible.

5 La courbe C<sub>1</sub> représente les caractéristiques de fréquence fondamentale d'un signal obtenu par une conversion linéaire classique.

Il apparaît que ce signal présente la même forme générale que celle du signal source représentée par la courbe Cx.

10 A l'inverse, la courbe C<sub>2</sub> représente les caractéristiques de fréquence fondamentale d'un signal converti à l'aide du procédé de l'invention tel que décrit en référence aux figures 2A et 2B.

Il transparaît de manière flagrante que la courbe de fréquence fondamentale du signal converti à l'aide du procédé de l'invention présente une forme générale très proche de la courbe de fréquence fondamentale cible Cy.

15 Sur la figure 4, on a représenté un schéma bloc fonctionnel d'un système de conversion de voix mettant en œuvre le procédé décrit en référence aux figures 2A et 2B.

20 Ce système utilise en entrée une base de données 120 d'échantillons vocaux prononcés par le locuteur source et une base de données 122 contenant au moins les mêmes échantillons vocaux prononcés par le locuteur cible.

Ces deux bases de données sont utilisées par un module 124 de détermination de fonctions de transformation de caractéristiques acoustiques du locuteur source en caractéristiques acoustiques du locuteur cible.

25 Ce module 124 est adapté pour la mise en œuvre des étapes 56 et 58 du procédé telles que décrites en référence à la figure 2 et permet donc la détermination d'une fonction de transformation de l'enveloppe spectrale des trames non voisées et d'une fonction de transformation conjointe de l'enveloppe spectrale et de la fréquence fondamentale des trames voisées.

30 De manière générale, on considère que le module 124 comporte une unité 126 de détermination de la fonction de transformation conjointe de l'enveloppe spectrale et de la fréquence fondamentale des trames voisées et une unité 128 de détermination de la fonction de transformation de l'enveloppe spectrale des trames non voisées.



Le système de conversion de voix reçoit en entrée un signal vocal 130 correspondant à un signal de parole prononcé par le locuteur source et destiné à être converti.

5 Le signal 130 est introduit dans un module 132 d'analyse du signal, mettant en œuvre, par exemple, une décomposition de type HNM permettant de dissocier des informations d'enveloppe spectrale du signal 130 sous la forme de coefficients cepstraux et des informations de fréquence fondamentale. Le module 132 délivre également des informations de phase et de fréquence maximale de voisement obtenues par l'application du modèle HNM.

10 Le module 132 met donc en œuvre l'étape 36 du procédé décrit précédemment et avantageusement l'étape 38.

Eventuellement cette analyse peut être faite au préalable et les informations sont stockées pour être utilisées ultérieurement.

15 Le système comporte ensuite un module 134 de séparation des trames voisées et des trames non voisées dans le signal vocal à convertir analysé.

Les trames voisées, séparées par le module 134, sont transmises à un module 136 de transformation adapté pour appliquer la fonction de transformation conjointe déterminée par l'unité 126.

20 Ainsi, le module 136 de transformation met en œuvre l'étape 104 décrite en référence à la figure 2B. Avantageusement, le module 136 met également en œuvre l'étape 108 de dénormalisation.

25 Les trames non voisées, séparées par le module 134, sont transmises à un module 138 de transformation adapté pour appliquer la fonction de transformation déterminée par l'unité 128 de manière à transformer les coefficients cepstraux des trames non voisées.

Ainsi, le module 138 de transformation des trames non voisées met en œuvre l'étape 106 décrite à la figure 2B.

30 Le système comporte également un module 140 de synthèse recevant en entrée, pour les trames voisées les informations d'enveloppe spectrale et de fréquence fondamentale transformées conjointement et les informations de phase et de fréquence maximale de voisement délivrées par le module 136. Le module 140 reçoit également les coefficients cepstraux des trames non voisées transformés et délivrés par le module 138.

Le module 140 met ainsi en œuvre l'étape 110 du procédé décrit en référence à la figure 2B et délivre un signal 150 correspondant au signal vocal 130 du locuteur source mais dont les caractéristiques d'enveloppe spectrale et de fréquence fondamentale ont été modifiées afin d'être similaires à celles du locuteur cible.

Le système décrit peut être mis en œuvre de diverses manières et notamment à l'aide des programmes informatiques adaptés et reliés à des moyens matériels d'acquisition sonores.

Dans le cadre de l'application du procédé de l'invention, tel que décrit en référence aux figures 1A et 1B, le système comporte dans le module 124, une unique unité de détermination d'une fonction de transformation conjointe de l'enveloppe spectrale et de la fréquence fondamentale.

Dans un tel mode de réalisation, les modules 134 de séparation et 138 d'application de la fonction de transformation des trames non voisées, ne sont pas nécessaires.

Le module 136 permet donc l'application de la seule fonction de transformation conjointe à toutes les trames du signal vocal à convertir et délivre les trames transformées au module 140 de synthèse.

De manière générale, le système est adapté pour la mise en œuvre de toutes les étapes des procédés décrits en référence aux figures 1 et 2.

Dans tous les cas, le système peut également être mis en œuvre sur des bases de données déterminées afin de former des bases de données de signaux convertis prêts à être utilisés.

Par exemple, l'analyse est faite en temps différé et les paramètres de l'analyse HNM sont mémorisés en vue d'une utilisation ultérieure lors des étapes 40 ou 100 par le module 134.

Enfin, en fonction de la complexité des signaux et de la qualité souhaitée, le procédé de l'invention et le système correspondant peuvent être mis en œuvre en temps réel.

Bien entendu d'autres modes de réalisation que ceux décrits peuvent être envisagés.

Notamment, les modèles HNM et GMM peuvent être remplacés par d'autres techniques et modèles connus de l'homme de l'art. Par exemple, l'analyse est réalisée à l'aide de techniques dites LPC (Linear Predictive

Coding), de modèles sinusoïdaux ou MBE (Multi Band Excited), les paramètres spectraux sont des paramètres dits LSF (Line Spectrum Frequencies), ou encore des paramètres liés aux formants ou à un signal glottique. En variante, le modèle GMM est remplacé par une quantification vectorielle (Fuzzy VQ.).

5           En variante, l'estimateur mis en œuvre lors de l'étape 30 est un critère de maximum a posteriori, dit "MAP" et correspondant à la réalisation du calcul de l'espérance uniquement pour le modèle représentant le mieux le couple de vecteurs source-cible.

10           Dans une autre variante, la détermination d'une fonction de transformation conjointe est réalisée à l'aide d'une technique dite des moindres carrés au lieu de l'estimation de la densité jointe décrite.

15           Dans cette variante, la détermination d'une fonction de transformation comprend la modélisation de la densité de probabilité des vecteurs source à l'aide d'un modèle GMM puis la détermination des paramètres du modèle à l'aide d'un algorithme EM. La modélisation prend ainsi en compte des segments de parole du locuteur source dont les correspondants prononcés par le locuteur cible ne sont pas disponibles.

20           La détermination comprend ensuite la minimisation d'un critère des moindres carrés entre paramètres cible et source pour obtenir la fonction de transformation. Il est à noter que l'estimateur de cette fonction s'exprime toujours de la même manière mais que les paramètres sont estimés différemment et que des données supplémentaires sont prises en compte.

**REVENDEICATIONS**

1. Procédé de conversion d'un signal vocal (130) prononcé par un locuteur source en un signal vocal converti (150) dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible, comprenant :

5                   - la détermination (1) d'au moins une fonction de transformation de caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches de celles du locuteur cible, à partir d'échantillons vocaux des locuteurs source et cible ; et

10                  - la transformation (2) de caractéristiques acoustiques du signal vocal à convertir (130) du locuteur source, par l'application de ladite au moins une fonction de transformation,

                    caractérisé en ce que ladite détermination (1) comprend la détermination (1; 56) d'une fonction de transformation conjointe de caractéristiques relatives à l'enveloppe spectrale et de caractéristiques relatives à  
15                  la fréquence fondamentale du locuteur source et en ce que ladite transformation (2) comprend l'application de ladite fonction de transformation conjointe.

2. Procédé selon la revendication 1, caractérisé en ce que ladite détermination (1 ; 56) d'une fonction de transformation conjointe comprend :

20                  - une étape (4X, 4Y) d'analyse des échantillons vocaux des locuteurs source et cible regroupés en trames pour obtenir, pour chaque trame d'échantillons d'un locuteur, des informations relatives à l'enveloppe spectrale et à la fréquence fondamentale ;

25                  - une étape (16X, 16Y ; 62X, 62Y) de concaténation des informations relatives à l'enveloppe spectrale et à la fréquence fondamentale pour chacun des locuteurs source et cible ;

                    - une étape (20 ; 70) de détermination d'un modèle représentant des caractéristiques acoustiques communes des échantillons vocaux du locuteur source et du locuteur cible ; et

30                  - une étape (30 ; 80) de détermination, à partir de ce modèle et des échantillons vocaux, de ladite fonction de transformation conjointe.

3. Procédé selon la revendication 2, caractérisé en ce que lesdites étapes d'analyse (4X,4Y) des échantillons vocaux des locuteurs source et cible sont adaptées pour délivrer lesdites informations relatives à l'enveloppe spectrale sous la forme de coefficients cepstraux.

4. Procédé selon la revendication 2 ou 3, caractérisé en ce que lesdites étapes (4X, 4Y) d'analyse comprennent chacune la modélisation des échantillons vocaux selon une somme d'un signal harmonique et d'un signal de bruit qui comprend :

- 5                   - une sous-étape (8X, 8Y) d'estimation de la fréquence fondamentale des échantillons vocaux ;
- une sous-étape (10X, 10Y) d'analyse synchronisée de chaque trame d'échantillons sur sa fréquence fondamentale ; et
- une sous-étape (12X, 12Y) d'estimation de paramètres
- 10 d'enveloppe spectrale de chaque trame d'échantillons.

5. Procédé selon l'une quelconque des revendications 2 à 4, caractérisé en ce que ladite étape (20 ; 70) de détermination d'un modèle correspond à la détermination d'un modèle de mélange de densités de probabilités gaussiennes.

- 15                   6. Procédé selon la revendication 5, caractérisé en ce que ladite étape de détermination (20 ; 70) d'un modèle comprend :

- une sous-étape (22, 72) de détermination d'un modèle correspondant à un mélange de densités de probabilités gaussiennes, et
- une sous-étape (24, 74) d'estimation des paramètres du mélange
- 20 de densités de probabilités gaussiennes à partir de l'estimation du maximum de vraisemblance entre les caractéristiques acoustiques des échantillons des locuteurs source et cible et le modèle.

- 7. Procédé selon l'une quelconque des revendications 2 à 6, caractérisé en ce que ladite détermination (1 ; 56) d'au moins une fonction de
- 25 transformation, comporte en outre une étape (14X, 14Y ; 60X, 60Y) de normalisation de la fréquence fondamentale des trames d'échantillons des locuteurs source et cible respectivement par rapport aux moyennes des fréquences fondamentales des échantillons analysés des locuteurs source et cible.

- 30                   8. Procédé selon l'une quelconque des revendications 2 à 7, caractérisé en ce qu'il comporte une étape (18 ; 50) d'alignement temporel des caractéristiques acoustiques du locuteur source avec les caractéristiques acoustiques du locuteur cible, cette étape (18 ; 50) étant réalisée avant ladite étape (20 ; 70) de détermination d'un modèle conjoint.

9. Procédé selon l'une quelconque des revendications 1 à 8, caractérisé en ce qu'il comporte une étape (54) de séparation dans les échantillons vocaux du locuteur source et du locuteur cible, des trames à caractère voisé et des trames à caractère non voisé, ladite détermination (56) d'une fonction de transformation conjointe des caractéristiques relatives à l'enveloppe spectrale et à la fréquence fondamentale étant réalisée uniquement à partir desdites trames voisées et le procédé comportant une détermination (58) d'une fonction de transformation des seules caractéristiques d'enveloppe spectrale uniquement à partir desdites trames non voisées.

10. Procédé selon l'une quelconque des revendications 1 à 8, caractérisé en ce que ladite détermination (1) d'au moins une fonction de transformation comprend uniquement ladite étape (1) de détermination d'une fonction de transformation conjointe.

11. Procédé selon l'une quelconque des revendications 1 à 10, caractérisé en ce que ladite détermination (1 ; 56) d'une fonction de transformation conjointe est réalisée à partir d'un estimateur de la réalisation des caractéristiques acoustiques du locuteur cible sachant les caractéristiques acoustiques du locuteur source.

12. Procédé selon la revendication 11, caractérisé en ce que ledit estimateur est formé de l'espérance conditionnelle de la réalisation des caractéristiques acoustiques du locuteur cible sachant la réalisation des caractéristiques acoustiques du locuteur source.

13. Procédé selon l'une quelconque des revendications 1 à 12, caractérisé en ce que ladite transformation (2) de caractéristiques acoustiques du signal vocal à convertir (130), comporte :

- une étape (36) d'analyse de ce signal vocal (130), regroupé en trames pour obtenir, pour chaque trame d'échantillons, des informations relatives à l'enveloppe spectrale et à la fréquence fondamentale ;

- une étape (38) de formatage des informations acoustiques relatives à l'enveloppe spectrale et à la fréquence fondamentale du signal vocal à convertir ; et

- une étape (40 ; 102) de transformation des informations acoustiques formatées du signal vocal à convertir (130) à l'aide de ladite fonction de transformation conjointe.

14. Procédé selon les revendications 9 et 13 prises ensemble, caractérisé en ce qu'il comporte une étape (100) de séparation, dans ledit signal vocal à convertir (130), des trames voisées et des trames non voisées, ladite étape de transformation comprenant :

5                   - une sous-étape (104) d'application de ladite fonction de transformation conjointe aux seules trames voisées dudit signal à convertir (130) ; et

                  - une sous-étape (106) d'application de ladite fonction de transformation des seules caractéristiques d'enveloppe spectrale auxdites trames non voisées dudit signal à convertir (130).

15                   15. Procédé selon les revendications 10 et 13 prises ensemble, caractérisé en ce que ladite étape de transformation comprend l'application de ladite fonction de transformation conjointe aux caractéristiques acoustiques de toutes les trames dudit signal vocal à convertir (130).

                  16. Procédé selon l'une quelconque des revendications 1 à 15, caractérisé en ce qu'il comporte en outre une étape (44 ; 110) de synthèse permettant de former un signal vocal converti (150) à partir des dites informations acoustiques transformées.

                  17. Système de conversion d'un signal vocal (130) prononcé par un locuteur source en un signal vocal converti (150) dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible, comprenant :

                  - des moyens (124) de détermination d'au moins une fonction de transformation des caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches du locuteur cible, à partir d'échantillons vocaux prononcés par les locuteurs source et cible : et

                  - des moyens (136, 138) de transformation des caractéristiques acoustiques du signal vocal à convertir (130) du locuteur source par l'application de ladite au moins une fonction de transformation,

                  caractérisé en ce que lesdits moyens (124) de détermination d'au moins une fonction de transformation, comprennent une unité (126) de détermination d'une fonction de transformation conjointe de caractéristiques relatives à l'enveloppe spectrale et de caractéristiques relatives à la fréquence fondamentale du locuteur source et en ce que lesdits moyens de transformation

comportent des moyens (136) d'application de ladite fonction de transformation conjointe.

18. Système selon la revendication 17, caractérisé en ce qu'il comporte en outre :

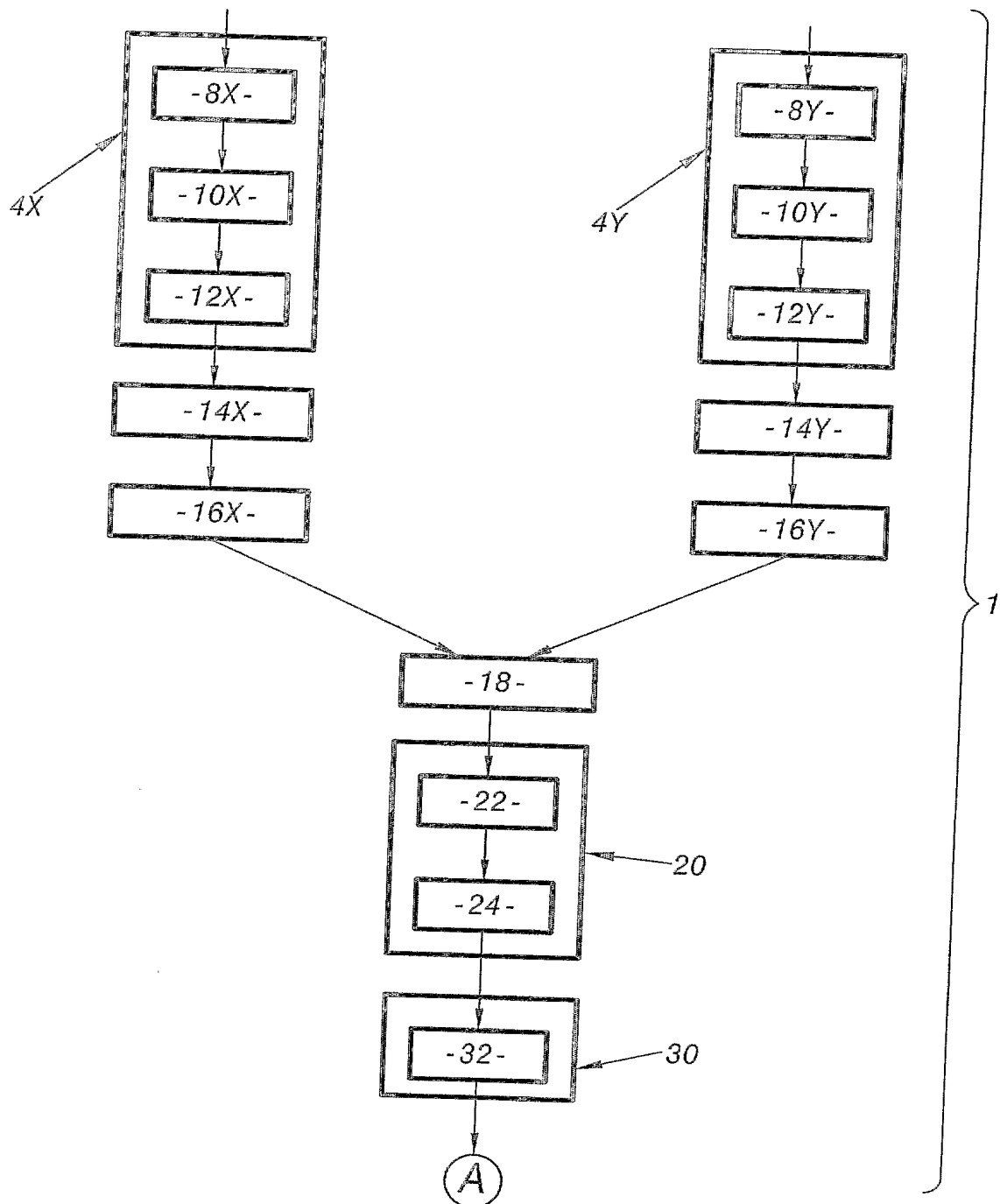
5                   - des moyens (132) d'analyse du signal vocal à convertir (130), adaptés pour délivrer en sortie des informations relatives à l'enveloppe spectrale et à la fréquence fondamentale du signal vocal à convertir (130) ; et

10                   - des moyens (140) de synthèse permettant de former un signal vocal converti à partir au moins desdites informations d'enveloppe spectrale et de fréquence fondamentale transformées simultanément.

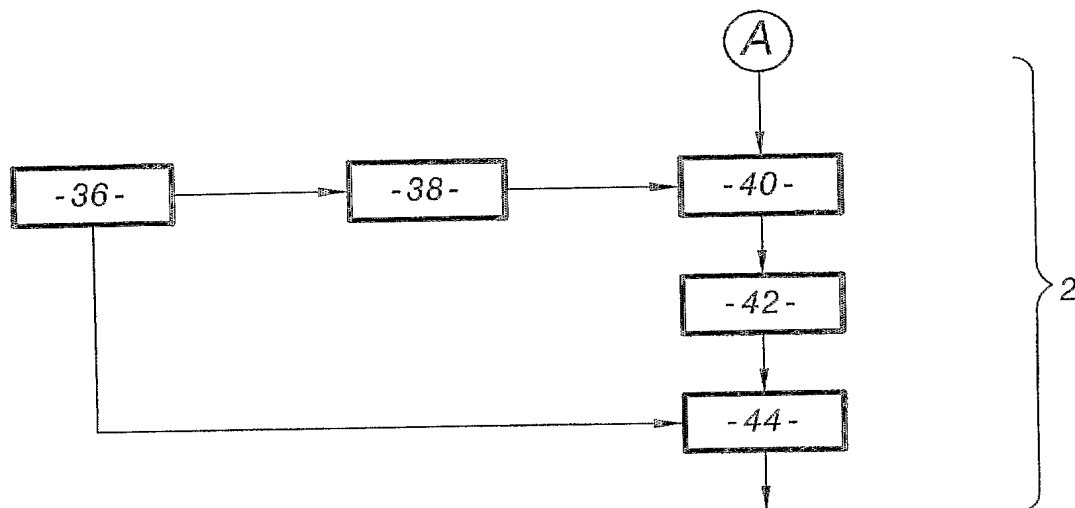
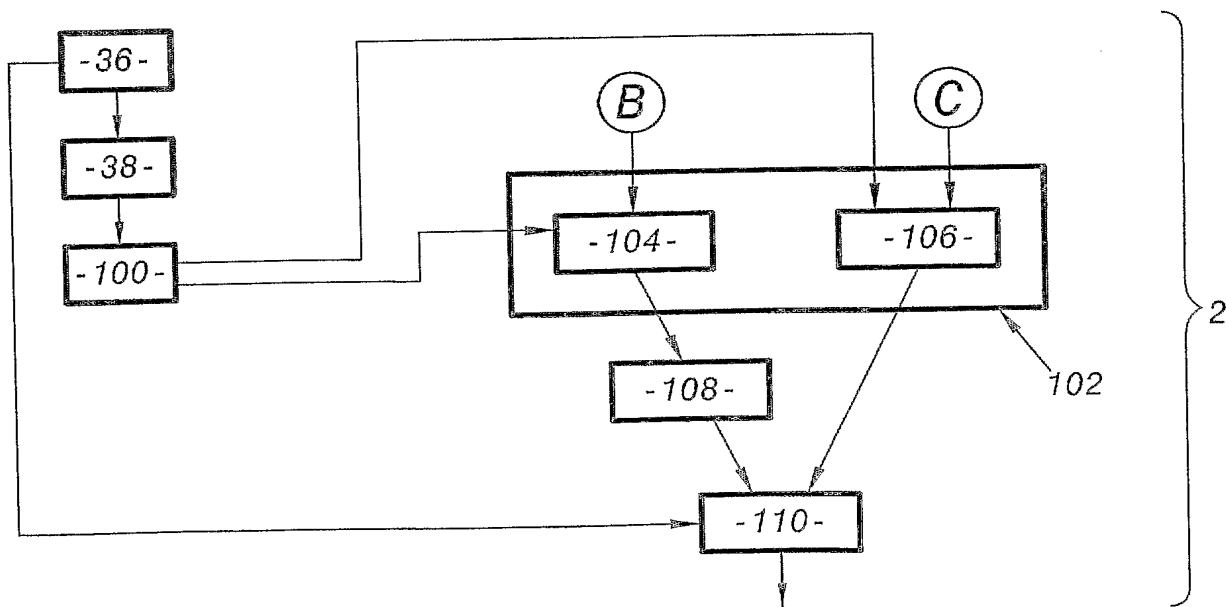
19. Système selon l'une quelconque des revendications 17 et 18, caractérisé en ce que lesdits moyens (124) de détermination d'au moins une fonction de transformation de caractéristiques acoustiques comportent en outre une unité (128) de détermination d'une fonction de transformation de l'enveloppe spectrale des trames non voisées, ladite unité (126) de détermination de la  
15 fonction de transformation conjointe étant adaptée pour la détermination de la fonction de transformation conjointe uniquement pour les trames voisées.



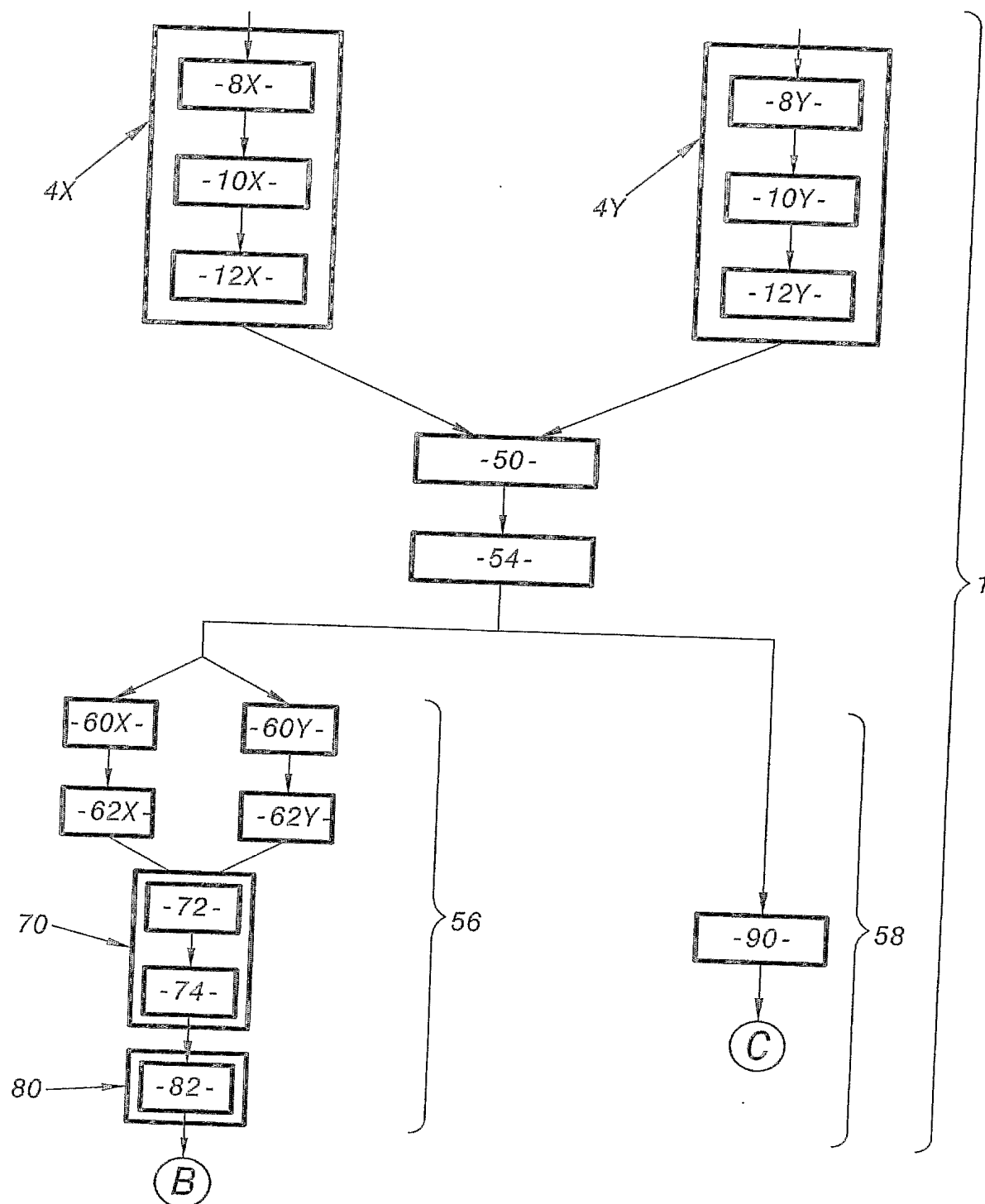
1/5

**FIG.1A**

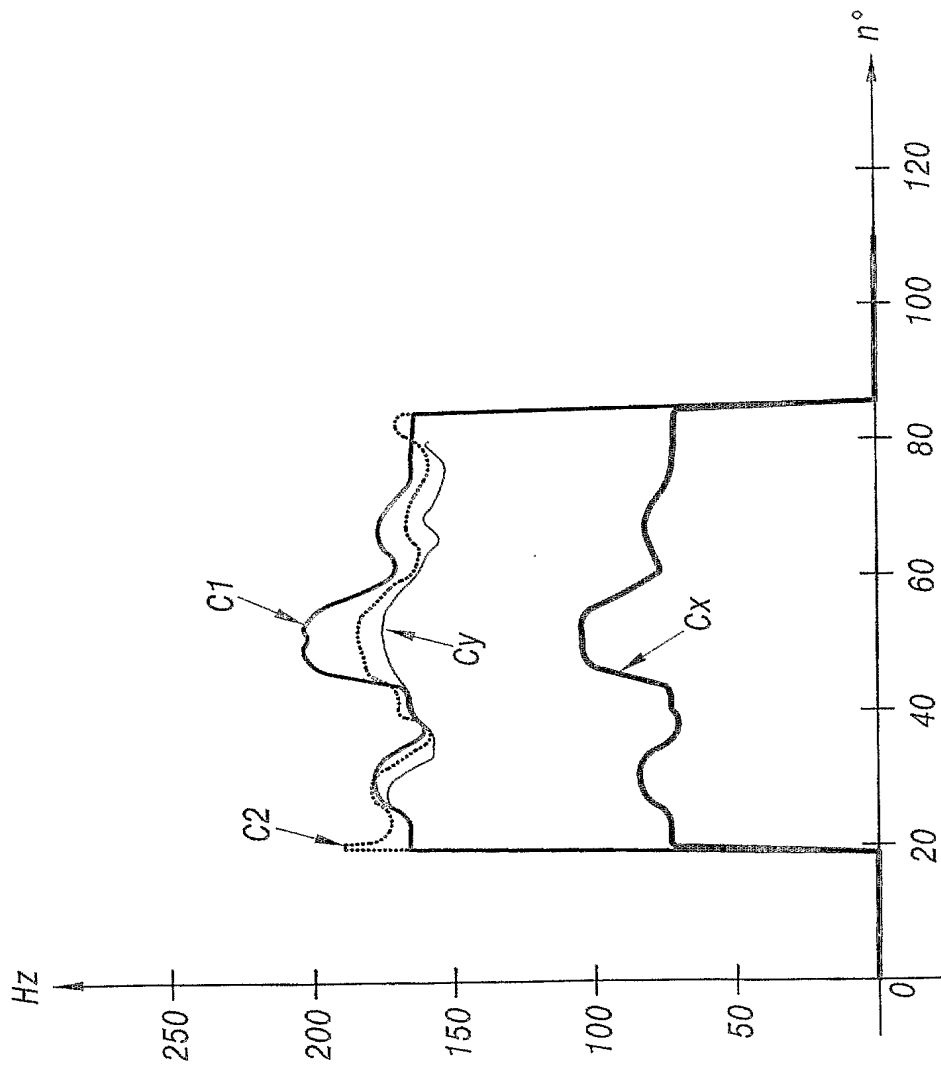
2/5

FIG. 1BFIG. 2B

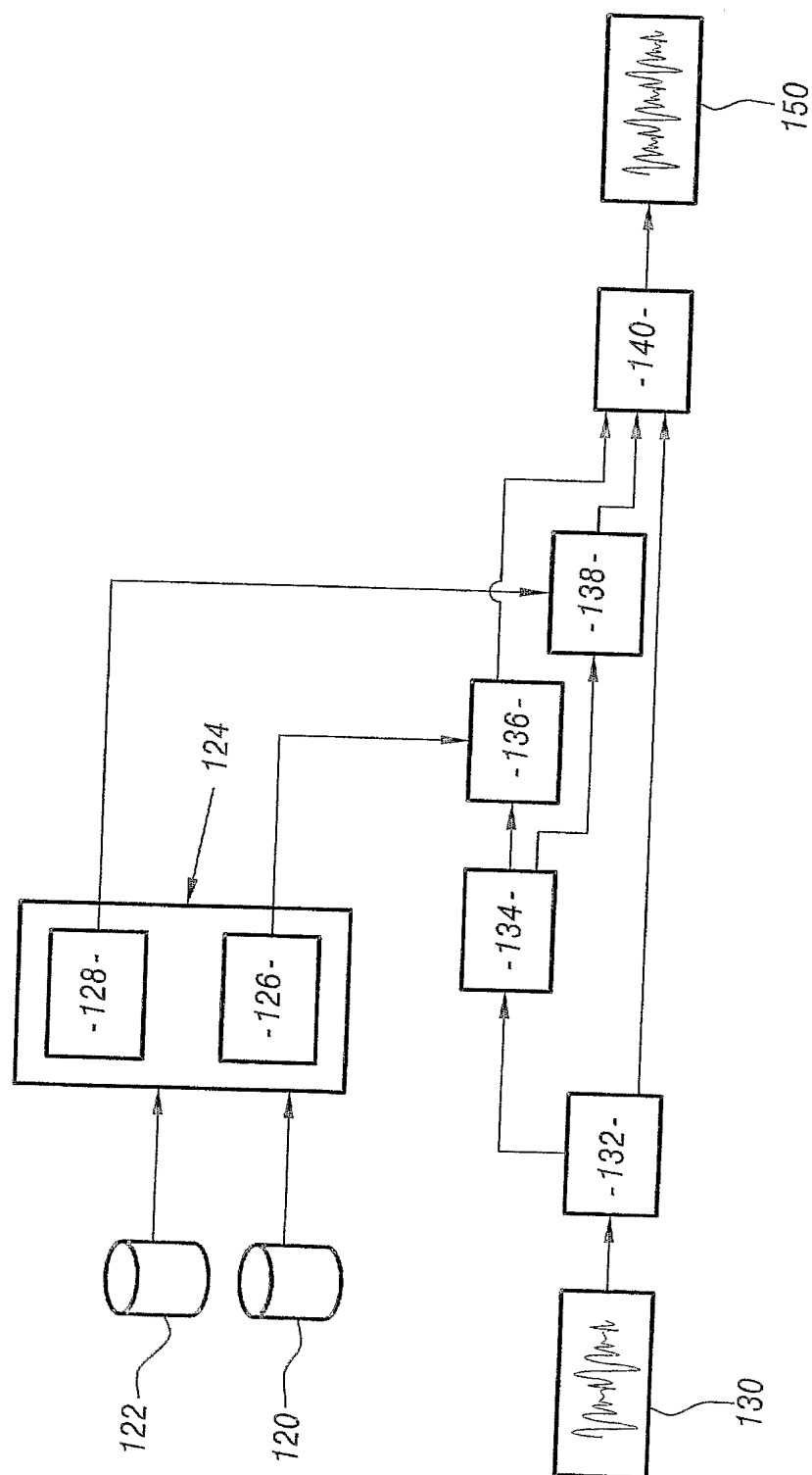
3/5

FIG. 2A

4/5

**FIG.3**

5/5

**FIG.4**



26 bis, rue de Saint Pétersbourg - 75800 Paris Cedex 08

Pour vous informer : INPI DIRECT

 0 825 83 85 87  
0,15 € TTC/mn

Télécopie : 33 (0)1 53 04 52 65

## BREVET D'INVENTION

## CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI



N° 11235\*03

DÉSIGNATION D'INVENTEUR(S) Page N° 1.. / 1..

(À fournir dans le cas où les demandeurs et les inventeurs ne sont pas les mêmes personnes)



Cet imprimé est à remplir lisiblement à l'encre noire

DB 113 @ W / 210103

Vos références pour ce dossier (facultatif)		BFF 04P0012
N° D'ENREGISTREMENT NATIONAL		0603603
TITRE DE L'INVENTION (200 caractères ou espaces maximum)		
Procédé et système améliorés de conversion d'un signal vocal.		
LE(S) DEMANDEUR(S) :		
FRANCE TELECOM		
DESIGNE(NT) EN TANT QU'INVENTEUR(S) :		
1	Nom	EN-NAJJARY
	Prénoms	Taoufik
Adresse	Rue	8, résidence Breiz
	Code postal et ville	1212131010 LANNION FRANCE
Société d'appartenance (facultatif)		
2	Nom	ROSEC
	Prénoms	Olivier
Adresse	Rue	29, rue André Gide
	Code postal et ville	1212131010 LANNION FRANCE
Société d'appartenance (facultatif)		
3	Nom	
	Prénoms	
Adresse	Rue	
	Code postal et ville	
Société d'appartenance (facultatif)		
S'il y a plus de trois inventeurs, utilisez plusieurs formulaires. Indiquez en haut à droite le N° de la page suivi du nombre de pages.		
DATE ET SIGNATURE(S) DU (DES) DEMANDEUR(S) OU DU MANDATAIRE (Nom et qualité du signataire)		
31 mars 2004 Ph. BLOT N° 98-0404		

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire. Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.

